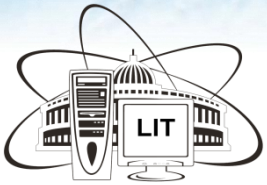


Tests of different MPI implementations in HPC/KVM cluster

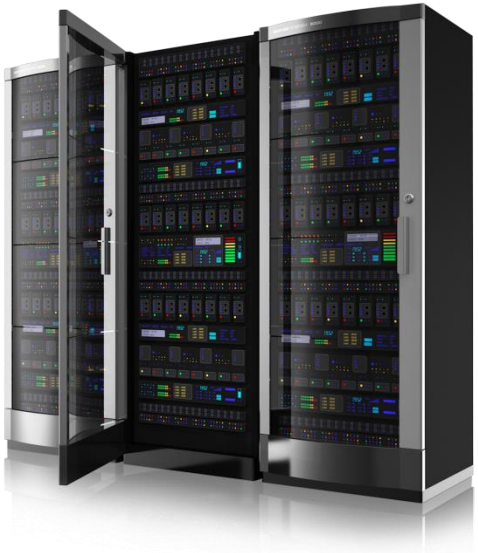
E.I Alexandrov, M.V Bashashin, D.V Belyakov, D.V Podgainy, O.I Streltsova, M.I Zuev

Joint Institute for Nuclear Research, Dubna

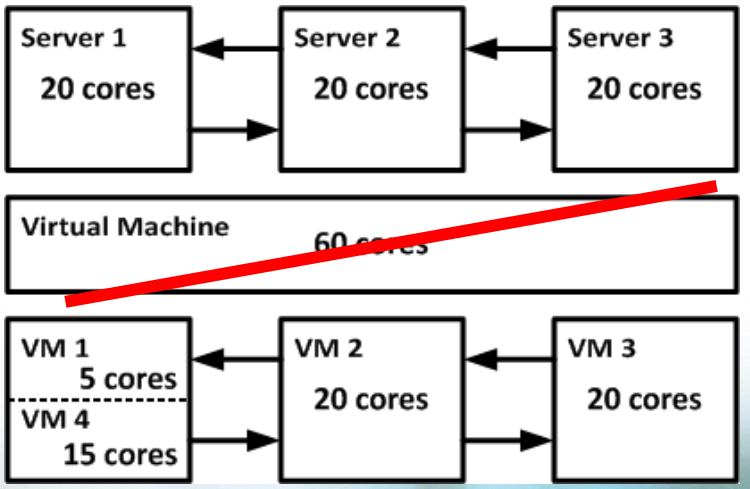
The work was financially supported by the RFBR grant No. 15-29-01217

Motivation: HPC or Cloud

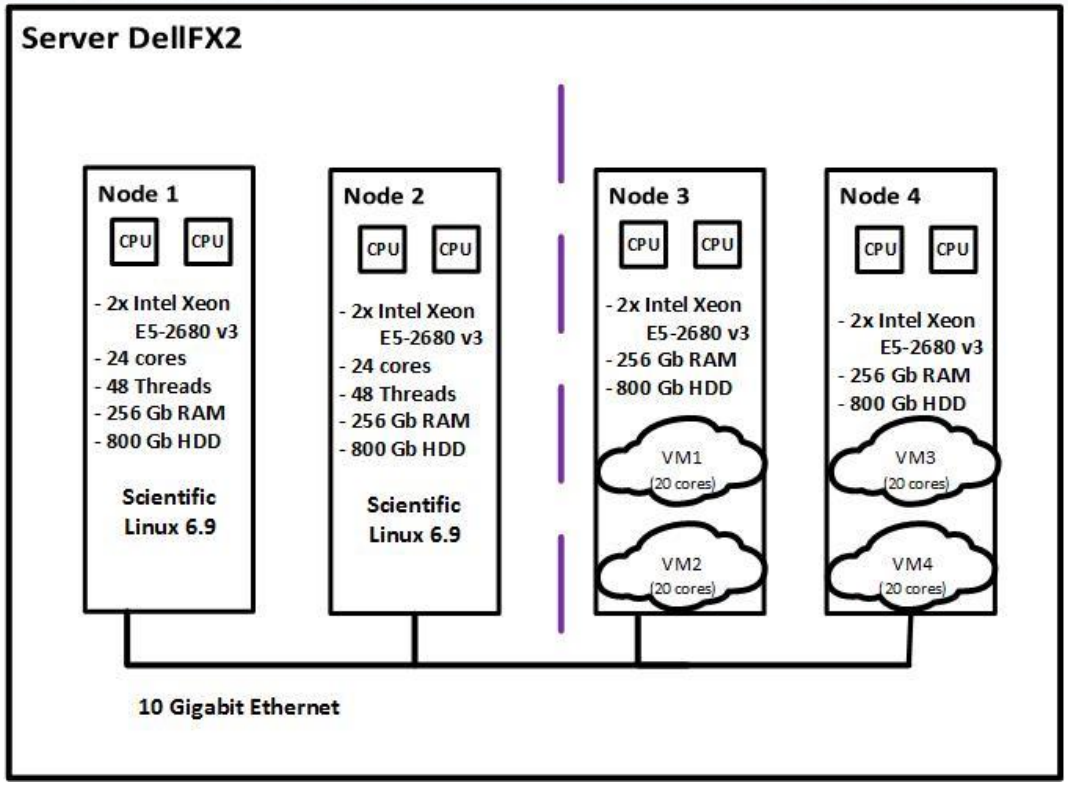
MPI



Hypervisors: KVM, OpenVZ, Xen



Testing area



Node3, Node4
24 Cores / 48 Threads
RAM 256 GB
CentOS-7
QEMU KVM 1.5.3

Node1, Node2
24 Cores /48 Threads
RAM 256 GB
Scientific Linux 6.9

VM1, VM2, VM3, VM4
20 Cores
RAM 20 GB
Scientific Linux 6.9

KVM Network

Network device

Drivers

- Intel PRO/1000 (E1000)
- Virtual Realtek 8139 (rtl8139)
- The para-virtualized network driver (virtio)

Network performance

Test tools: iperf3 Version: 3.0.12

Description: a tool for active measurements of the maximum achievable bandwidth on IP networks

Hosts	Driver	Performance	CPU	Comments
Node1-Node2		9.41 Gbit/s	15%	Between real node
VM1-VM3	e1000	1.80-1.93 GBit/s	100%	
VM1-VM3	rtl8139	258 MBit/s	100%	
VM1-VM3	virtio	9.41 Gbit/s	25%	
VM1-VM2	virtio	24 GBit/s or 33 GBit/s	82%	Two VM on one blade. Result is stable but depend on usage core (real or virtual)

MPI Benchmark

Test tools: **IMB-MPI1** Version: 2017.2.050

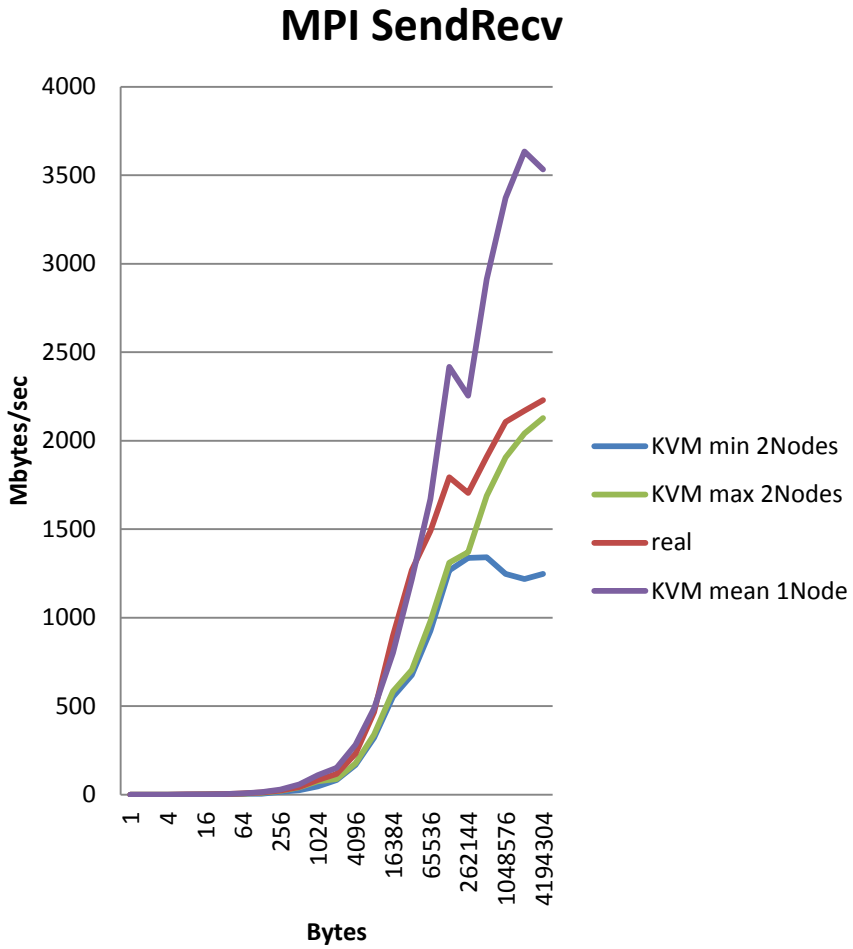
Description: The Intel® MPI Benchmarks perform a set of MPI performance measurements for point-to-point and global communication operations for a range of message sizes. **IMB-MPI1** - benchmarks for MPI-1 functions.

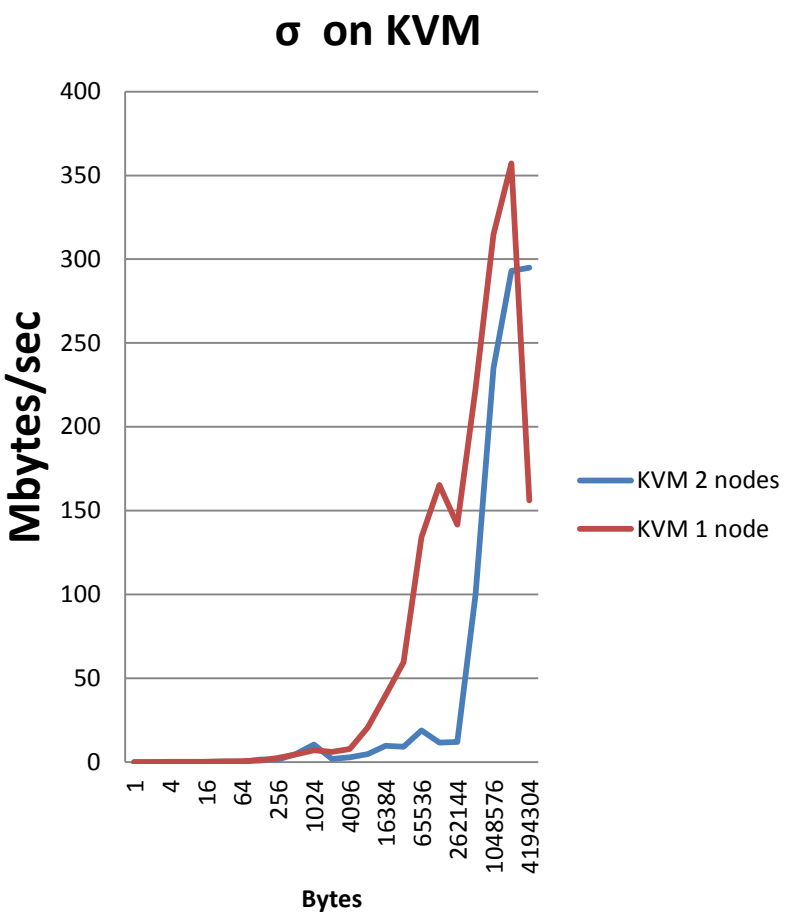
MPI-1 functions:

- PingPong
- PingPing
- Sendrecv
- Exchange
- Allreduce
- Reduce
- Reduce_scatter
- Allgather
- Allgatherv
- Gather
- Gatherv
- Scatter
- Scatterv
- Alltoall
- Alltoallv
- Bcast
- Barrier

These results are shown in the presentation

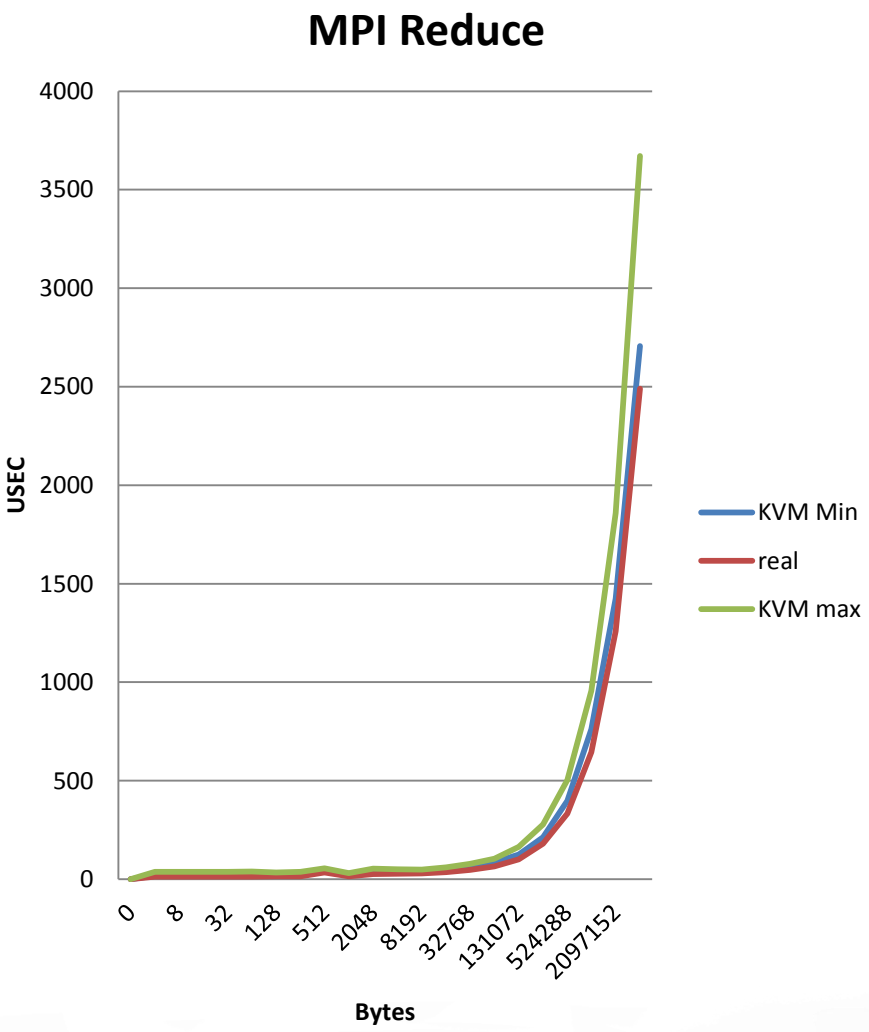
MPI Benchmark: Sendrecv

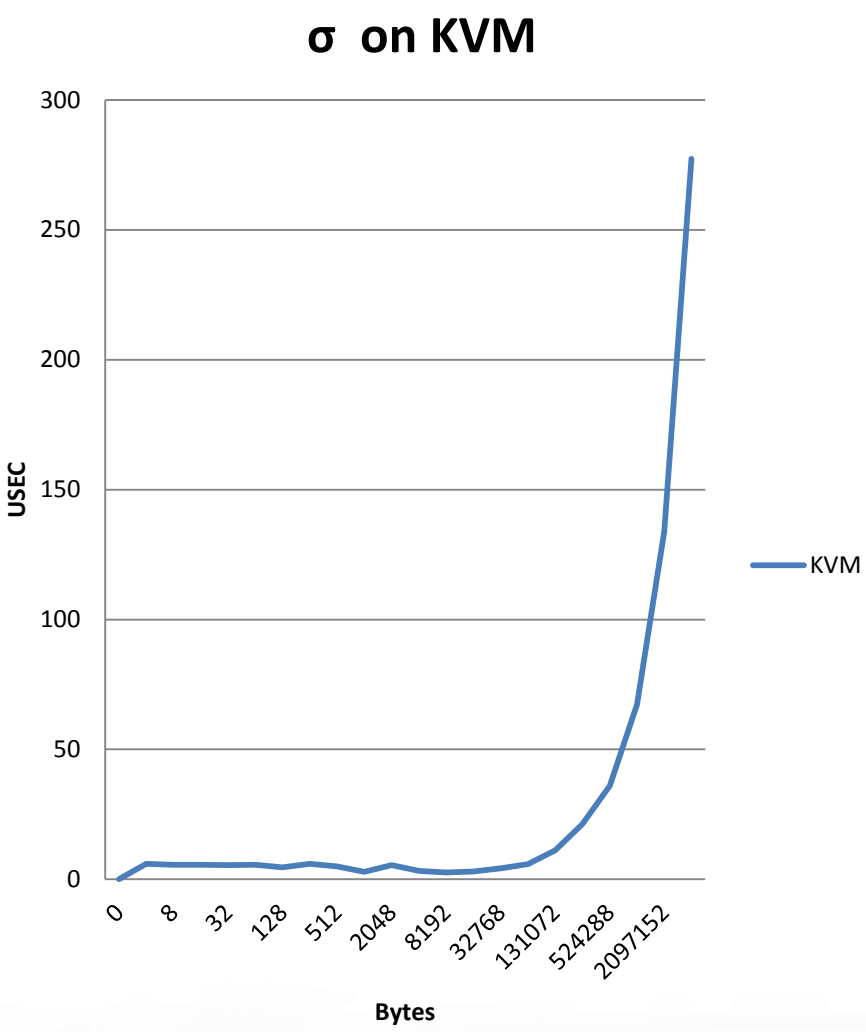




Number of tests is 10. Each test has several repetitions (from 10 for big number of bytes to 1000 for small number of bytes).

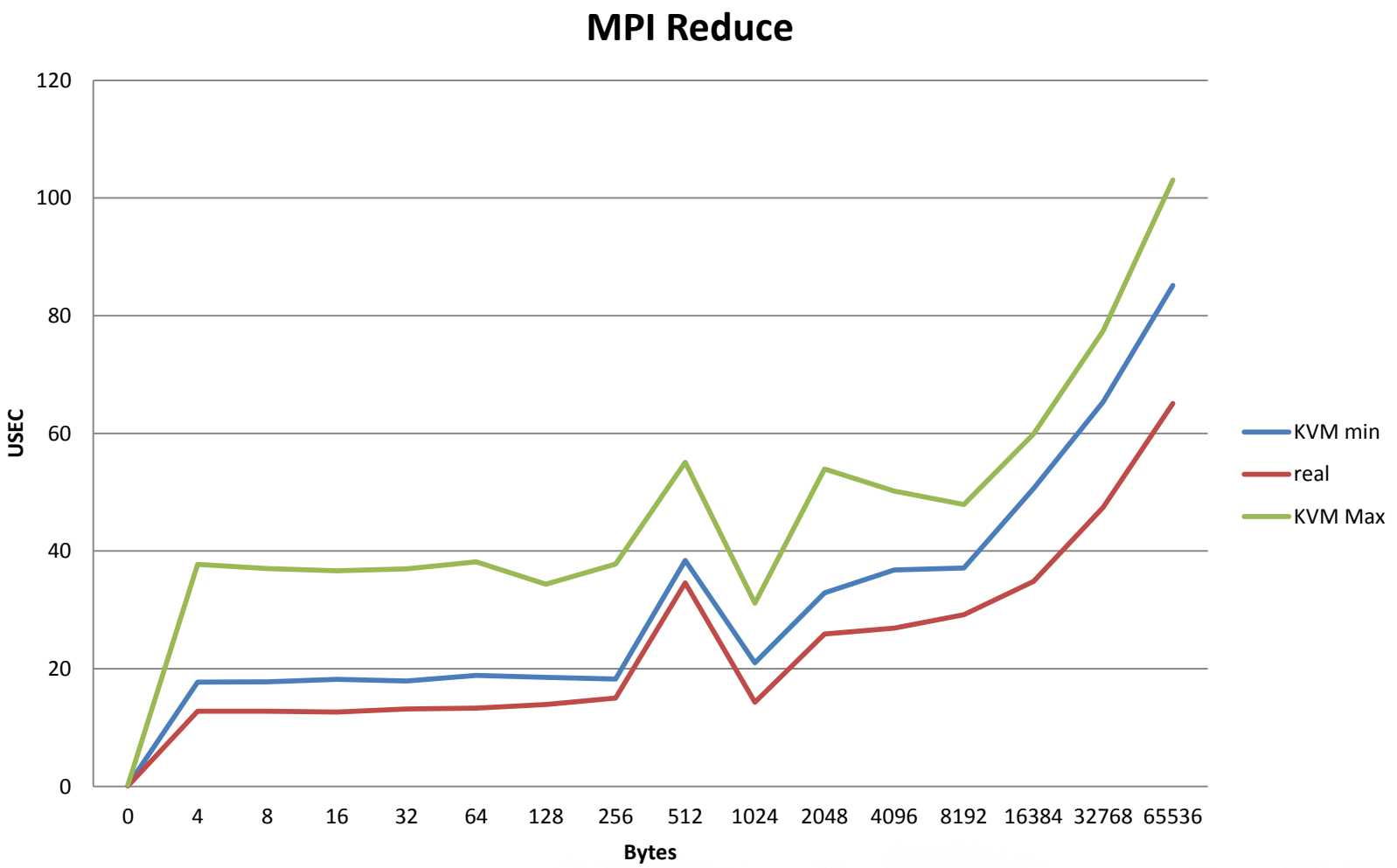
MPI Benchmark: Reduce





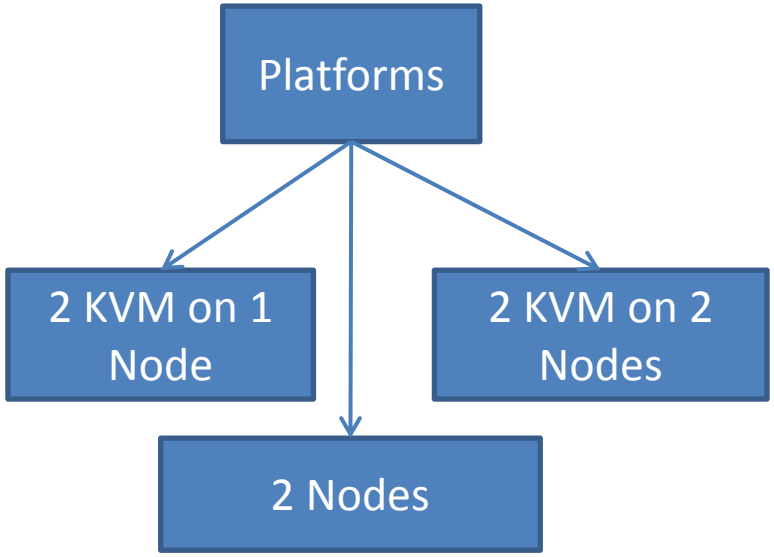
Number of tests is 10. Each test has several repetitions (from 10 for big number of bytes to 1000 for small number of bytes).

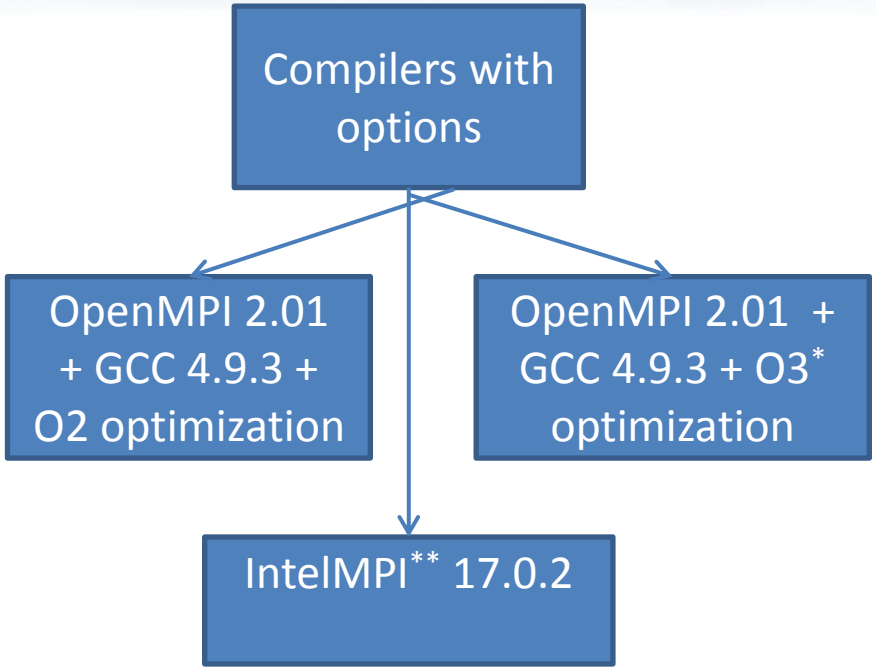
MPI Benchmark: Reduce

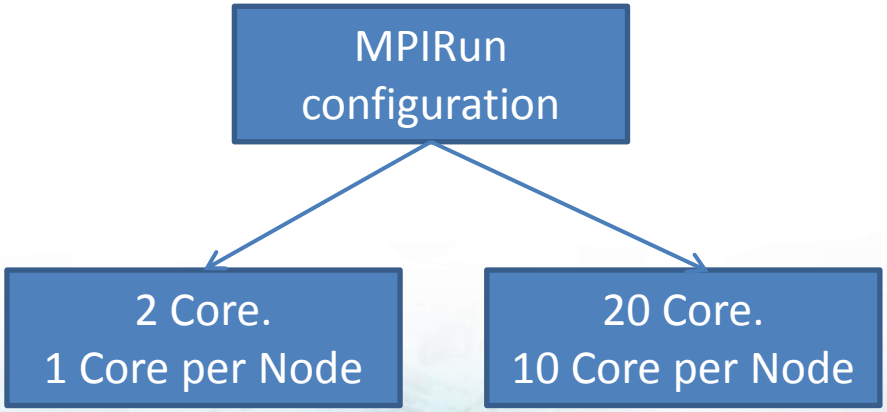


Number of tests is 10. Each test has several repetitions (from 10 for big number of bytes to 1000 for small number of bytes).

Environment of tests on real programs







*Not all programs can compile
**Intel O2 and O3 has like results

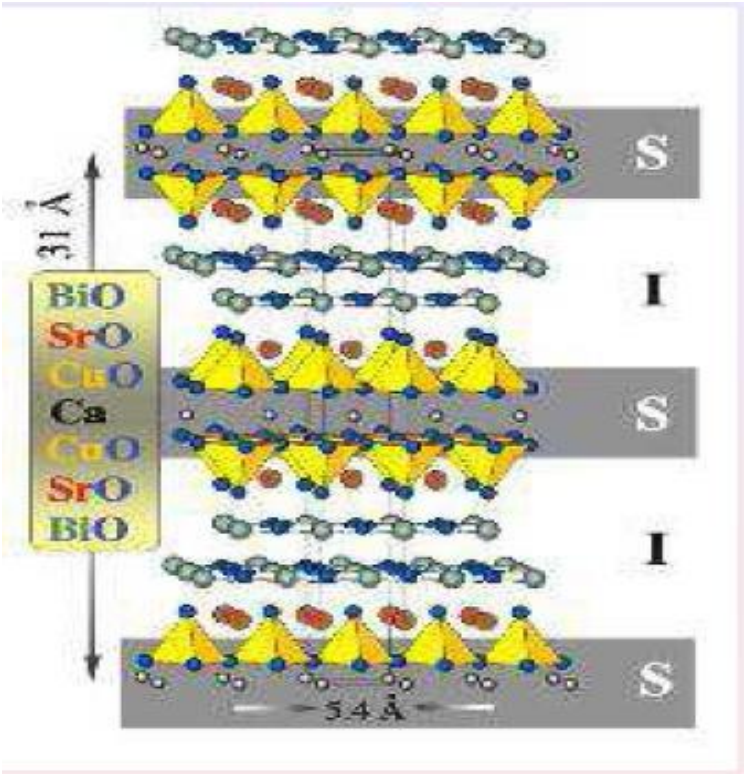
Long Josephson Junctions (LJJ)

Parallel computer program* is used for simulation of superconducting processes in LJJ system in dependence on the parameters of capacitive and inductive coupling.

LJJ system consists of superconducting layers with intermediate dielectric (insulator) layers of length *L.*

Size of MPI send receive data is 10 Double values.

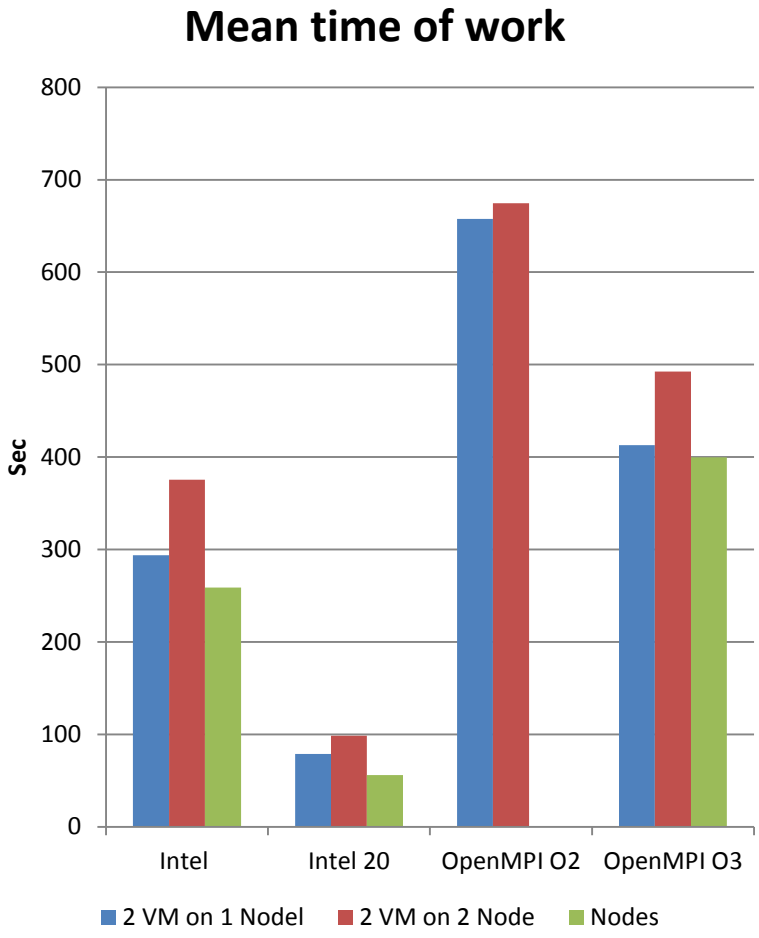
Structure of the natural crystal $Bi_2Cr_2CaCu_2O_8$ with superconducting properties

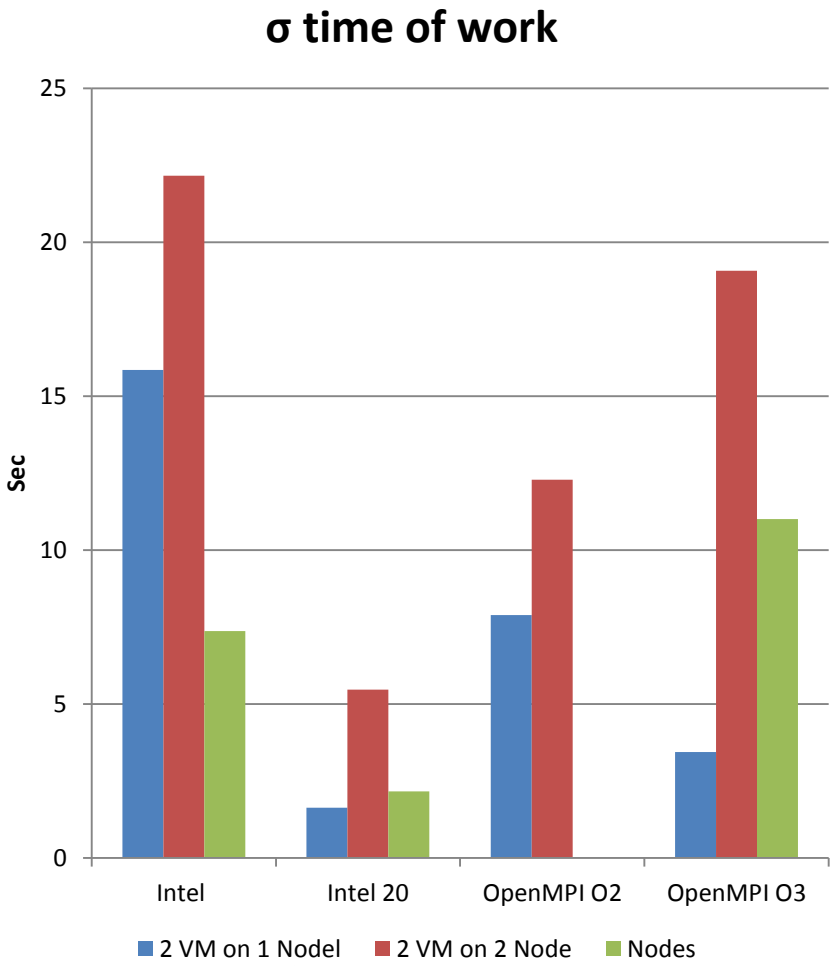


S – superconducting layers,
I – insulator (dielectric) layers

*Atanasova P., Bashashin M.V., Rahmonov I.R., Shukrinov Yu.M., Volohova A.V., Zemlyanaya E.V. Numerical approach and parallel implementation for computer simulation of stacked long Josephson Junctions // Computer Research and Modeling. – P. 8, № 4. – 2016. – Pp. 593-604.

Long Josephson Junctions test



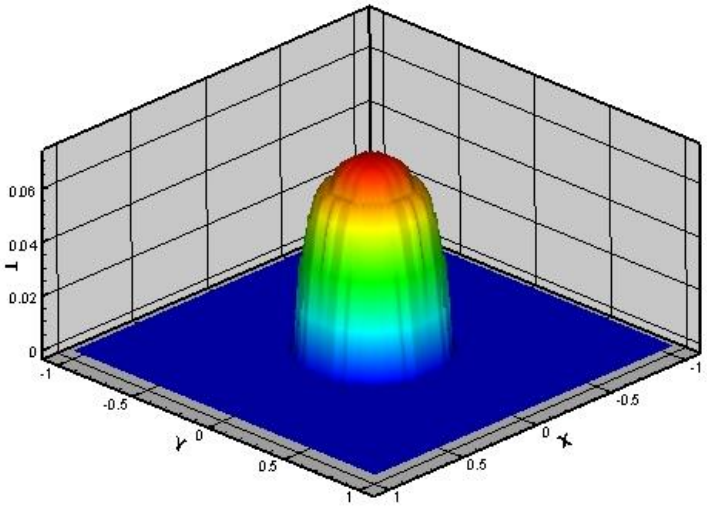


Number of tests is 10.

GIMM_FPEIVE

GIMM_FPEIVE* is the parallel computer program is used for 2D modeling of thermal processes in materials irradiated with ion beams.

Result of work:

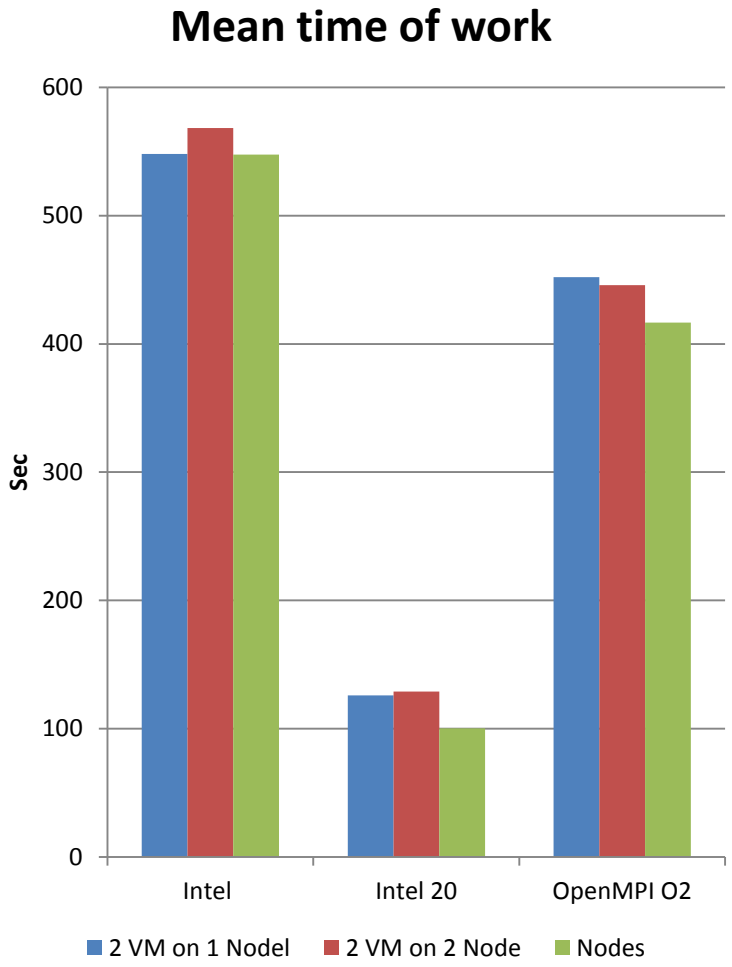


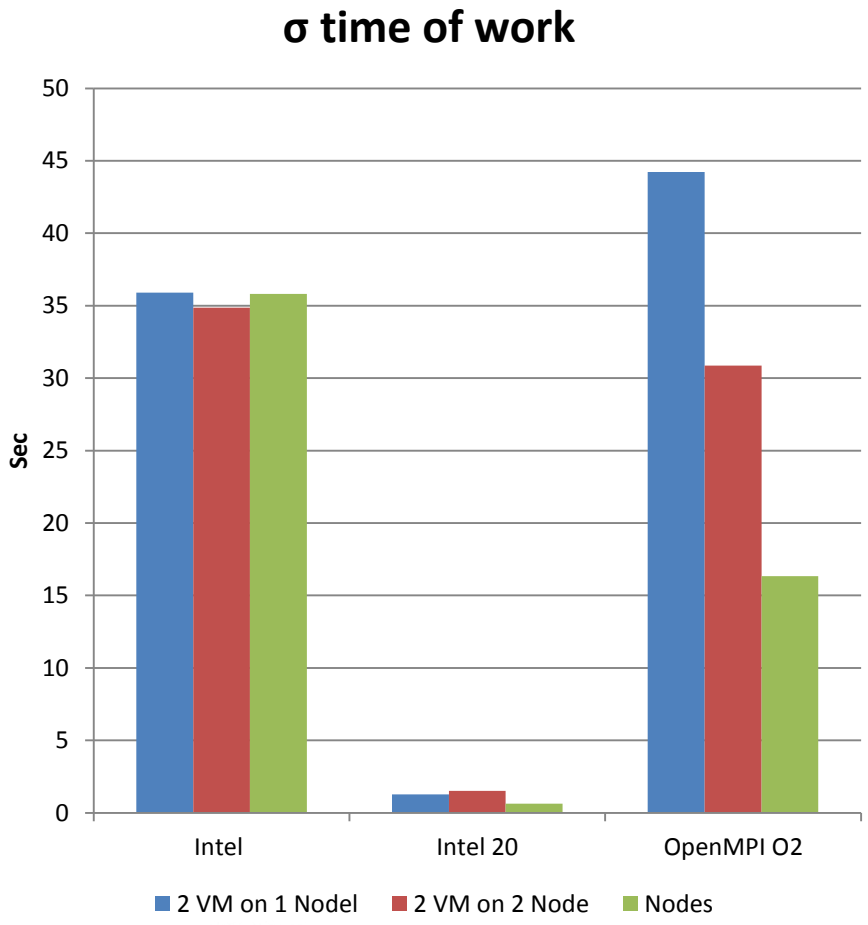
Input parameters:

Ion:	AU
Target:	Ni
Energy:	700 MeV
NumberZ**:	5001
NumberR:	513

- *E. Alexandrov, I. Amirkhanov et all. Principles of Software Construction for Simulation of Physical Processes on Hybrid Computing Systems (on the Example of GIMM FPEIP Complex) // RUDN Journal of Mathematics, Information Sciences and Physics, №2 – 2014. – Pp 197-205.
- **This axis use for parallelization.

GIMM_FPEIVE test





Number of tests is 10.

Conclusion

- Network driver is critical for VM
- Network between VMs on different node work the same as between real node
- Network between VMs on 1 node work faster that between real node, but required more CPU
- MPI operations between VMs on different node work slowly that between real node (difference depend on data size)
- MPI operations between VMs on 1 node work faster that between real node (difference depend on data size)
- Parallel computer program is used for simulation of superconducting processes (small size of MPI data) work faster on real node
  - Slowly about 10% for VMs on 1 Node
  - Slowly about 20-40% for VMs on 2 Node
- GIMM_FPEIVE (average size of MPI data) work like
  - Slowly about 0.1 – 8% for VMs on 1 Node
  - Slowly about 4 – 7% for VMs on 2 Nodes
- Big number of MPI process give degradation of speed on VMs